Exponential Concentration Inequality for a Rényi-$\alpha$ Divergence Estimator

Shashank Singh $^1$  Barnabás Póczos $^1$

June 22, 2014

$^1$Carnegie Mellon University, Pittsburgh, PA, USA
Given $\alpha \in [0, 1) \cup (1, \infty)$, estimate the Rényi-$\alpha$ divergence

$$D_\alpha(p\|q) = \frac{1}{\alpha - 1} \log \int_X p^\alpha(x) q^{1-\alpha}(x) \, dx,$$

between two unknown, continuous, nonparametric probability densities $p$ and $q$ over $X = [0, 1]^d$, using $n$ samples from each density.
Contribution

- plug-in estimator of Rényi-$\alpha$ divergence based on kernel density estimation
- bound bias of the estimator
- prove a concentration inequality
- simple proof-of-concept experiment
Motivation

- ‘distributional’ machine learning algorithms
  - finite-dimensional feature vectors $\rightarrow$ distribution features
Motivation

- ‘distributional’ machine learning algorithms
  - finite-dimensional feature vectors → distribution features
- KL-divergence, entropy, and mutual information special cases
  - applications to feature selection, clustering, ICA, etc.
Motivation

- ‘distributional’ machine learning algorithms
  - finite-dimensional feature vectors → distribution features
- KL-divergence, entropy, and mutual information special cases
  - applications to feature selection, clustering, ICA, etc.
- with concentration inequality:
  - can simultaneously bound error of multiple estimates (e.g., forest density estimation)
  - can derive hypothesis test for independence
Related Work

- Few known rates
- No estimators have concentration inequalities or other results describing their distribution
Smoothness (Hölder) Condition

Same assumptions on $p$ and $q$. 

\[ \beta \text{-Hölder condition on } p, \quad L > 0, \quad \ell := \lfloor \beta \rfloor (\text{so } \beta - 1 \leq \ell < \beta) \]

All $\ell$-order (mixed) partial derivatives of $p$ and $q$ exist and

\[ \sup_{x \neq y \in X} |D_{\vec{i}}^p(x) - D_{\vec{i}}^p(y)| \leq L. \]
Smoothness (Hölder) Condition

Same assumptions on $p$ and $q$.

$\beta$-Hölder condition on $p$:

- $\beta, L > 0$, $\ell := \lfloor \beta \rfloor$ (so $\beta - 1 \leq \ell < \beta$)

All $\ell$-order (mixed) partial derivatives of $p$ and $q$ exist and

$$\sup_{x \neq y \in \mathcal{X}, |i| = \ell} \frac{|D^i p(x) - D^i p(y)|}{\|x - y\|^{\beta - \ell}_r} \leq L.$$
There exist known $\kappa_1, \kappa_2 \in \mathbb{R}$ such that, $\forall x \in \mathcal{X}$,

$$0 < \kappa_1 \leq p(x), q(x) \leq \kappa_2 < +\infty.$$ 

- *Existence* of $\kappa_2$ is trivial, but our estimator requires it to be *known* beforehand.
- Assuming $\kappa_1$ for $q$ is natural (to ensure $D_\alpha(p\|q) < +\infty$).
- $\kappa_1$ for $p$ is technical, and can be weakened/eliminated in certain cases.
- Reason for working on finite measure domain $\mathcal{X} = [0, 1]^d$. 

Boundary Condition

All derivatives of $p$ vanish at the boundary; i.e.,

$$\sup_{1 \leq |i| \leq \ell} |D^i p(x)| \to 0$$

as

$$\text{dist}(x, \partial \mathcal{X}) \to 0.$$
All derivatives of $p$ vanish at the boundary; i.e.,

$$\sup_{1 \leq |i| \leq \ell} |D^\vec{i} p(x)| \to 0$$

as

$$\text{dist}(x, \partial \mathcal{X}) \to 0.$$ 

Strong assumption, but needed to eliminate boundary bias.
Kernel Assumptions

\[ K : \mathbb{R} \to \mathbb{R} \] with support in \([-1, 1]\) and satisfies

\[
\int_{-1}^{1} K(u) \, du = 1 \quad \text{and} \quad \int_{-1}^{1} u^j K(u) \, du = 0, \quad \forall j \in \{1, \cdots, \ell\}.
\]
1. Mirror data $x^1, \ldots, x^n$ across all subsets of edges of $\mathcal{X}$
1. Mirror data $x^1, \cdots, x^n$ across all subsets of edges of $X$.
2. Using a bandwidth $h$ and product kernel $K^d$, compute kernel density estimate (KDE) $\tilde{p}$ from resulting $3^d n$ data points.
Mirrored Kernel Density Estimate

1. Mirror data $x^1, \cdots, x^n$ across all subsets of edges of $\mathcal{X}$
2. Using a bandwidth $h$ and product kernel $K^d$, compute kernel density estimate (KDE) $\tilde{p}$ from resulting $3^d n$ data points
   - Removes boundary bias because we assume derivatives of $p$ vanish near $\partial \mathcal{X}$. 
Rényi-$\alpha$ Divergence Estimator

1. Clip mirrored KDE below by $\kappa_1$ and above by $\kappa_2$

   i.e., $\hat{p}(x) = \min\{\kappa_2, \max\{\kappa_1, \tilde{p}(x)\}\}$.

2. Compute $\hat{q}$ by the same process

3. Plug $\hat{p}, \hat{q}$ into $D_{\alpha}$:

   $$D_{\alpha}(\hat{p} \parallel \hat{q}) = \frac{1}{\alpha - 1} \log \int_{X} \hat{p}^\alpha(x) \hat{q}^{1-\alpha}(x) \, dx.$$
Bounds

- **Bias Bound**: \( \exists C_B \in \mathbb{R} \) such that

\[
|\mathbb{E} D_\alpha(\hat{p} \parallel \hat{q}) - D_\alpha(p \parallel q)| \leq C_B \left( h^\beta + h^{2 \beta} + \frac{1}{nh^d} \right).
\]

- **Concentration Inequality (‘Variance’ Bound)**: \( \exists C_V \in \mathbb{R} \) such that, \( \forall \varepsilon > 0 \),

\[
P \left( |D_\alpha(\hat{p} \parallel \hat{q}) - \mathbb{E} D_\alpha(\hat{p} \parallel \hat{q})| > \varepsilon \right) \leq 2 \exp \left( -C_V^2 \varepsilon^2 n \right).
\]
Bias Bound

\[ |\mathbb{E} D_\alpha(\hat{p}\|\hat{q}) - D_\alpha(p\|q)| \leq C_B \left( h^\beta + h^{2\beta} + \frac{1}{nh^d} \right) . \]

Proof Sketch:

1. Main step is to analyze boundary bias of mirrored KDE:
   \[ \int_X (\mathbb{E} \hat{p}(x) - p(x))^2 \, dx \leq C_b h^{2\beta} . \]

2. Rest is a technical blend of standard proof techniques
Concentration Inequality

\[ P(\left| D_{\alpha}(\hat{p} \parallel q) - \mathbb{E}D_{\alpha}(\hat{p} \parallel q) \right| > \varepsilon) \leq 2 \exp \left( -C_{V}^{2} \varepsilon^{2} n \right) \]

Proof Sketch:

By McDiarmid’s Inequality, suffices to bound change in estimator by \( C_{V} / n \) when resampling one data point.

By Mean Value Theorem, change is proportional to integrated change in mirrored KDE.

By construction of KDE, this is proportional to \( 2 \| K \|_{d,1} / n \).


Concentration Inequality

\[ \mathbb{P} \left( |D_\alpha(\hat{p} \parallel \hat{q}) - \mathbb{E}D_\alpha(\hat{p} \parallel \hat{q})| > \varepsilon \right) \leq 2 \exp \left( -C_V^2 \varepsilon^2 n \right) \]

Proof Sketch:

- By McDiarmid’s Inequality, suffices to bound change in estimator by \( C_V/n \) when resampling one data point.
Concentration Inequality

\[ P \left( | D_\alpha(\hat{p} \parallel \hat{q}) - \mathbb{E} D_\alpha(\hat{p} \parallel \hat{q}) | > \varepsilon \right) \leq 2 \exp \left( -C_V^2 \varepsilon^2 n \right) \]

Proof Sketch:

- By McDiarmid’s Inequality, suffices to bound change in estimator by \( C_V/n \) when resampling one data point.
- By Mean Value Theorem, change is proportional to integrated change in mirrored KDE.
Concentration Inequality

\[ \mathbb{P}( |D_\alpha(\hat{p} \parallel \hat{q}) - \mathbb{E}D_\alpha(\hat{p} \parallel \hat{q})| > \varepsilon ) \leq 2 \exp\left(-C_V^2\varepsilon^2 n\right) \]

Proof Sketch:

- By McDiarmid’s Inequality, suffices to bound change in estimator by $C_V/n$ when resampling one data point.
- By Mean Value Theorem, change is proportional to integrated change in mirrored KDE.
- By construction of KDE, this is proportional to $2\|K\|_1^d/n$. 
Consequences

- Can bound variance by integrating concentration inequality:

\[ \nabla \mathbb{V}[D_\alpha(\hat{p} \| \hat{q})] \leq C_V^2 n^{-1}. \]
Consequences

• Can bound variance by integrating concentration inequality:
  \[ \mathbb{V}[D_\alpha(\hat{p} \parallel \hat{q})] \leq C_V^2 n^{-1}. \]

• Choose bandwidth \( h \) to minimize bias bound asymptotically:
  \( h \asymp n^{-\frac{1}{\beta+d}} \). Then,
  - Bias is \( O \left( n^{-\frac{\beta}{\beta+d}} \right) \)
  - MSE is \( O \left( n^{-\frac{2\beta}{\beta+d}} + n^{-1} \right) \)
  - parametric rate \( O(n^{-1}) \) if \( \beta \geq d \) and slower \( O \left( n^{-\frac{2\beta}{\beta+d}} \right) \) else
Estimated divergence between two Gaussians in $\mathbb{R}^3$.

**Figure**: Log-log plot of empirical MSE alongside theoretical bound. Error bars indicate standard deviation of estimator from 100 trials.
Present an estimator of Rényi-$\alpha$ Divergence
Summary

- Present an estimator of Rényi-$\alpha$ Divergence
- Prove $O\left(n^{-\frac{\beta}{\beta+d}}\right)$ bias bound
Summary

- Present an estimator of Rényi-$\alpha$ Divergence
- Prove $O\left(n^{-\frac{\beta}{\beta+d}}\right)$ bias bound
- Prove exponential concentration of estimator
Summary

- Present an estimator of Rényi-$\alpha$ Divergence
- Prove $O\left(n^{-\frac{\beta}{\beta+d}}\right)$ bias bound
- Prove exponential concentration of estimator
- Experimentally verify results
Future Work

1. Study role of dimension $d$
2. Prove concentration inequality for estimator of conditional quantities
   - e.g., Conditional Mutual Information:

   $$I_\alpha(X; Y|Z) = \int_Z D_\alpha(P(X, Y|Z)\|P(X|Z)P(Y|Z)) \, dP(Z)$$

   - hypothesis test for conditional independence
Thanks!
References

- **f-Divergence estimation:**

- **k-NN estimation:**

- **Lower bounds for single-density functional estimation:**

- **Distributional Machine Learning:**